1I1-R-4-2

RGB・深度センサを用いた複数人数会話の三次元記録環境の構築

Construction of 3-Dimensional Recording Environments for Multi-party Conversation with RGB-Depth Sensors

矢野 正治[*1] 大本 義正[*1] 西田 豊明[*1]
Masaharu Yano Yoshimasa Ohmoto Toyoaki Nishida

[*1]京都大学 情報学研究科 知能情報学専攻
Dept. of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

For analyses of multi-party conversation with participants' movements, we have developed a system to record and display conversation scenes in the form of 3D model data. The system displays 3D models and estimates skeleton positions by integrating participants' movements recorded by multiple RGB-Depth cameras. It complements occlusions by participants or objects, and reconstructs both 3D models of participants and that of a conversation environment in an about 3m square field.

1. はじめに

非言語情報を伴う会話の分析は，会話音声とともに人の振る舞いをビデオ映像として同時に記録し，閲覧してアノテーションを付与することでなされる．非言語情報を分析することで，会話内容だけでは分からない，会話者の興味度や発話意思などが推定できるようになる．過去の研究で，二者対面や，席に座って位置が固定された会話映像の記録はなされてきたが，複数人数で移動を伴うような会話では常に各人の動きが完全に映る様に撮影することは難しい．

本研究では，ビデオ映像のみでなく深度測定により深度映像を記録するセンサカメラを用いることで，移動を伴う会話を三次元的に記録・再現することを目的としている．三次元記録により，通常は配置できない位置からの映像や，会話参加者から見える映像をなどを再現することが可能となる．また，センサにより参加者の骨格位置推定を行い，参加者の位置と向きに応じた正面・対面映像の自動再現も目的とする．可搬性のあるセンサを用いることで，収録システムが固定された環境以外の複数の場所での撮影を可能にすることも目指す．

本研究では，これら会話シーンの三次元記録と姿勢推定を行うシステムを構築する．システムは，会話シーンの人の動きを三次元的に記録する部分と，会話環境全体を静的な三次元モデルとして復元する部分から成り，環境内で人のみが移動するような会話の場全体を三次元的に記録することを可能にする．

2. 関連研究

2.1 会話収録

ビデオ映像を記録する会話コーパスとしては，AMI[McCowan 05] が挙げられる。これらの会話は，多人数でミーティングを行っている場面を記録して会話コーパスとしているものであるが，会話参加者はそれぞれ所定の位置に座っており，会話環境内の自由な移動を伴う会話ではない。会話参加者の会話中の位置と向きが決まっている場合には，その前後にそれぞれビデオカメラを配置することで，会話中の人の動きと，人の動きの対象物との両方をほとんど観察することができるが，人の移動を伴うような会話ではそのようなカメラ配置を行うことは難しい．

連絡先: 矢野正治，京都大学 情報学研究科 知能情報学専攻，yano@ii.ist.i.kyoto-u.ac.jp

これらの研究に対して，本研究では複数人の会話参加者が会話環境の中である程度の範囲を自由に動き回るような会話シーンに対して，自動的に正面映像や背後からの映像を再現することを，会話参加者にセンサを直接取り付けることなく実現することを目的としている．

2.2 多人数会話シーンや周辺環境の三次元記録

人の動きを三次元データとして復元しようという研究は，コンピュータビジョンの分野で長年なされている [Moeslund 01]。これらの研究によって，一人の人間が周りに物がない環境で動きを行う場合の三次元形状復元は実用に達しているが，複数人の動きについては人同士の位置関係によって遮蔽が起こったりすることもあるため依然として難しい。

本研究では，複数人数の動きに対しても三次元モデルの復元や骨格推定を可能とすることを目的としている．また，環境の三次元復元についても，深度センサを用いることでできるだけ正確に，会話シーンの三次元モデルとの位置関係の対応がとれるように復元することを目的としている．

3. 会話の三次元記録のためのシステム

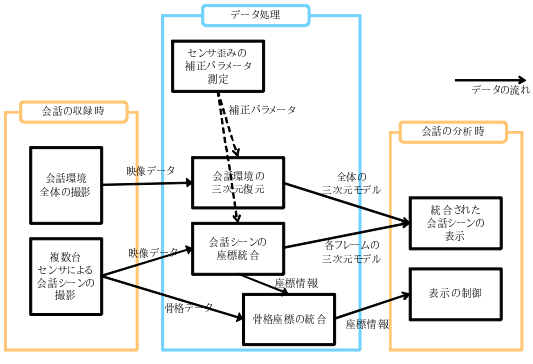

図 1: システム全体の構成

本研究では，複数人数の会話参加者の動きを任意の視点から再現し，その映像を用いて会話者同士や環境に対する非言語行動を分析するために，会話を三次元的に記録するシステムを構築する．そのシステムにより，各時点での人の腕や胴，頭部

の位置や形などが会話場の周囲や内部の視点から欠落無く閲覧できること，また，人の動きと会話の周囲の環境の位置関係が同様に閲覧できることを目指す．

システムの全体の構成は，図 1 の様になっている．左側はデータ収録部，中央はデータ処理の部分，右はデータ利用の部分を表している．データ収録に関して，会話シーン，会話環境のどちらに対しても同種の RGB・深度センサを用いるが，会話シーンに対しては複数の個所に固定した位置から撮影するのに対し，会話環境については一つのセンサを移動させながら撮影する．収録システムは特定の環境に固定されたものではないので，収録する環境に合わせて配置を調整することができる．

データ処理部に関しては主に四つの要素がある．ひとつは，複数台センサから得られる映像から三次元モデルを作成し，座標統合によりシーンの三次元復元を行う部分である．統合を容易に可能にすることで，特定の環境以外でも会話を収録できるようになる．二つ目は，複数台センサから得られる骨格情報を統合する部分である．参加者に直接センサを取り付けることなく動きをデータとして取得することができる．これにより固定点からだけでなく自動で主観的や対面的な映像を復元できる．三つ目は，センサを移動させながら環境全体を撮影したデータに対して，各フレームのセンサの座標を推定し，RGB・深度映像から三次元モデルを復元する，会話環境の三次元復元の部分である．会話シーンの復元だけでは会話環境全体をカバーしきれないので，主観映像に欠落が生じないように環境を別個に復元する必要がある．最後に，復元を正確に行うための，センサによる歪みの影響を測定し，補正を行うというセンサ歪みの補正パラメータ測定部分がある．

4. システムによる会話収録

4.1 収録対象とする会話シーン

三次元記録を行う会話として，低い高さの移動不可能な台や，壁などにかかっている物など指差しや視線の対象物があり得る環境内で，2〜4 人の会話参加者が 3m 四方程度の範囲を移動しながら会話を行う場面を対象としている．例えば，博物品の展示の様な，環境内に対象物が固定されていて，その対象物について相談や説明をする場面を想定しており，周囲さまざまな位置から対象物を見て回る会話での動きや見ている向きを映像として再現する．

会話内で起こる人の動きとしては，位置の移動，指差し，ハンドジェスチャといった，体や腕を動かす動作のうち，他者と接触しない様なものを三次元記録できる対象としている．会話環境の広さについて，Kinect の画角が垂直方向で 45 度程度なので，斜め上から撮影するようにしたうえで，撮影範囲の周囲に 1m 程度の余裕が必要となる．また，画像特徴量により会話環境の三次元復元を行うので，ある程度模様のある背景の場所である必要がある．

4.2 ハードウェア

RGB・深度センサとして，Microsoft 社の Kinect™ を用いる．Kinect から PC にデータを取り込むためのドライバ・ミドルウェアとして OpenNI を用いる．OpenNI を用いると，Kinect から RGB 映像・深度映像・人領域映像および人のトラッキング ID と骨格位置情報が取得できる．RGB 映像については 1280 × 1024 の画素数で 15 フレーム毎秒，または 640 × 480 の画素数で 30 フレーム毎秒の画像が取得できる．深度映像は，640 × 480 の画素数で 30 フレーム毎秒の画像が取得できる．実際のデータの解像度は赤外線パターンの粒度によって決まるので，データの解像度は画像の解像度と比べると数分の一となっている．また，実際には深度の計算には誤差が含まれているため，精度の高いデータの得られる距離は 4m 程度である．

5. RGB・深度映像からの会話の三次元復元

本章では，会話シーンの三次元記録，会話環境の三次元復元に関してそれぞれ実際のシステムの実装の詳細を述べる．システム全体の構成としては図 1 の様になっており，この章ではその中央部分の領域について詳しく述べる．

5.1 Kinect データの補正と補完

Kinect から得られる映像は，レンズの歪みを伴う．レンズの歪みを解消するため，まず，RGB 映像に対して OpenCV のチェスボードによるカメラキャリブレーション関数を用いて補正を行う．

深度値についても歪みがあり，レンズの主点上のピクセルでは精度が高いが，そこから離れたピクセルでは誤差が生じる．測定の精度を上げるために，ピクセルごとに誤差の大きさを推定して補正を行う．平面を様々な距離から撮影し，深度値の傾きを求めて正しい距離をピクセル毎に推測し，最小二乗法で測定値から期待値への一次変換をピクセル毎に求める．

また，現状の Kinect では，深度映像取得のためのパターン照射用の赤外線の波長を変更することができないため，同一表面上に対して Kinect を向けると赤外線パターンが互いに干渉し，深度値の欠落する領域が生じる．これを解消するため，[Maimone 11] の手法を用い，欠落の補完を行う．

5.2 会話シーンの三次元記録

5.2.1 複数台 Kinect の座標統合

測定対象に応じて Kinect の配置を容易に変更できるようにするため，環境内で Kinect がどの位置に配置されているかを Kinect から得られるデータを用いて計算する．まず複数方向から一人の人を撮影して得られるそれぞれの骨格位置を対応点として各 Kinect の座標を求めた後，深度映像を用いて近傍点探索により対応点を求めて細かい補正を行う．対応点から座標の変換の計算は，クォータニオンによる最小二乗法を用いる．

5.2.2 会話参加者の姿勢推定

会話参加者の姿勢推定は，OpenNI により計算された骨格位置推定の値を用いる．まず，骨格位置を統合するために，骨格点間の距離を用いて，単一 Kinect での一意 ID の付与と，Kinect 間での ID の統合を行う．次に，統合された ID ごとに，骨格点の統合を行う．人の映り方によっては推定誤差が生じているので，その影響を出来るだけ抑えるように統合する．範囲内外の移動や遮蔽により一部しか映っていない骨格点については，映っている骨格点の数を重みにすることでそれら誤りの大きい位置推定の影響を抑えることができる．向きに関しては，正面方向である場合の重みを大きくすることで横向き時の推定誤りの影響を抑えられる．近すぎて全身が映っていない場合は，垂直方向の分散を重みにすることで除外する．

また，骨格点推定で左右が入れ替わって推定されることもある．左右の入れ替えは，動きの速い場合などトラッキングの継続している状態でも起こり得るので，フレーム単位で，入れ替えの有無でフレーム間の各骨格点の距離の近いほうを採用して修正する．

5.3 会話環境の三次元復元

5.3.1 移動する Kinect の座標推定

会話シーンの三次元記録に際し，背景部分となる会話環境は一部は収録されるが，全体を映すのは難しく，また遮蔽も

生じる．そのため，Kinect を移動させながらデータを記録し，SLAM の手法で Kinect の座標推定と映像からの三次元復元を行う

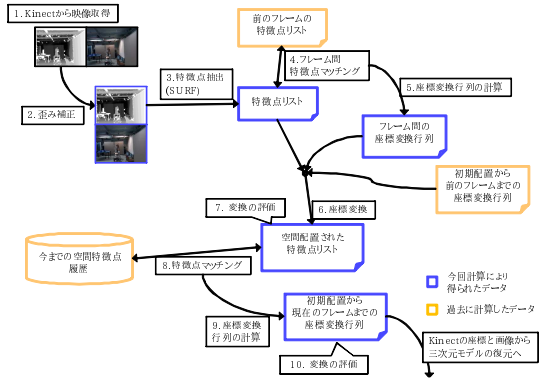

図 2: 得られた RGB・深度映像から Kinect の位置を推定する際の一フレームの画像ごとの処理の流れ

図 2 に，Kinect の三次元座標推定の流れのうち，各フレームに対する処理を示す．各フレームに対して，以前のフレームの画像特徴点とマッチングを行って座標推定を行っている．各フレームの座標が求まれば，深度映像から各ピクセルの絶対座標が求まり，三次元復元が可能となる．フレームのぶれや飛びなどを考慮して，以下の処理を行うようにしている．

まず，環境の復元をリアルタイムで行う必要はないため，あらかじめ映像を記録しておいて，一フレームの処理が完了するごとに取り出し，歪み補正処理を施す．RGB 画像をグレースケール化し，SURF により特徴点を求める．特徴点は特徴量ベクトルを持つので，二点間の非類似度が求められる．

フレーム間で各特徴点間の非類似度を求め，特徴点のマッチングを行う．特徴点の対応からクォータニオンの最小二乗法で座標変換を行う．最小二乗法では特徴点間の距離と非類似度，深度映像の値，面の方線の向きから求めた重みを用いる．また LMedS の手法により，誤ったマッチングの影響を抑えて座標変換を行う．得られた座標変換による位置変動の誤差が大きい場合，変換を破棄する．

フレーム間だけでは変換誤差が蓄積するので，過去の特徴点をある程度記録しておいて，絶対座標に移した今回のフレームと特徴点マッチングを行う．非類似度の小さな対応が多いので，複数対応をとり，様々な選択で行列を求める．フレーム間の変換行列と同じように，変換の誤差が一定より大きい場合に，フレームの飛びやぶれなどとみなして破棄し，三次元モデルの復元を行わないようにする．

5.3.2 環境の三次元モデル化

5.3.1 で各フレームの時点における Kinect の座標が推定できたので，各フレームの RGB 画像，深度画像と組み合わせて三次元点群を絶対座標にプロットすることで三次元空間の復元を行うことができる．本研究では滑らかなサーフェスの作成は行わず，座標を格子状に区切り固定サイズのボクセルごとに特徴点から色を与えることで環境の三次元モデルとした．

5.4 環境の三次元モデルと会話シーンの三次元モデルの統合

環境復元の際の座標系を，会話シーンの三次元モデルの座標系に合わせる形で座標推定を行うことで，会話シーンと環境の三次元モデルの座標系を合わせることができる．具体的には，会話シーンの映像から特徴点を求め，会話シーンの座標系上にマッピングし，それから環境の初期座標を求めることによって行う．

6. 実験と評価

6.1 環境の三次元復元と歪み補正

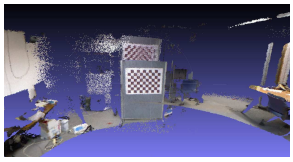

図 3: 歪み補正なしで一周分を三次元復元した際のずれ

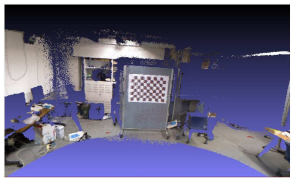

図 4: 歪み補正後の三次元復元の例

環境復元に際してどのくらいの精度で復元できるのかを評価するために，部屋の中心で Kinect を回転させて周囲を一周撮影したデータから三次元復元を行った．部屋のサイズはおよそ 8m×6m×2.7m 程度である．

歪み補正なしで復元した場合の結果を図 3 に示す．ずれのため，チェスボードが上下に重複して表示されている．一方，同じデータに対して歪み補正の処理を各フレームに行い，三次元復元をしたものを図 4 に示す．図 3 と比べて，一周した際の位置のずれがほとんど軽減されている．

図 5: 環境全体の三次元モデルの復元結果

さらに環境内を走査したデータを追加して，環境内のほとんどを復元した結果を図 5 に示す．中央の床部分や物体の陰になっている部分は再現できていないが，データの追加を繰り返すことで全体が三次元モデルとして復元できると考えられる．

6.2 会話シーンの三次元復元

6.2.1 複数 Kinect の座標系の統合

まず，複数台の Kinect で人ひとりを撮影して三次元復元を行い，どの程度再現できているかを調べた．

図 6 の様に，ひとりの人が指差しをしている場面を，三台の Kinect を用いて撮影した．得られた画像データに対し歪み補正の処理を施したのち，座標系を統合して，三次元復元を行った．結果は図 7 の様になっている．十分な解像度がないため，人の顔などは詳細が分かる様には再現されていないが，指差しや頭の向きといった人の動きは再現されている．特に，Kinect は斜め前方においてあるが，三次元モデルを真横の向きから描画した場合でも，腕の向きが分かるようになっている．

6.2.2 複数人数の会話シーンの姿勢推定

三人が会話参加者として立ち話をしている様子を，環境内に四方からと斜め上からの 5 台の Kinect を中心に向かって撮

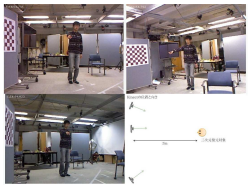

図 6: 指差しジェスチャ撮影時のカメラの配置

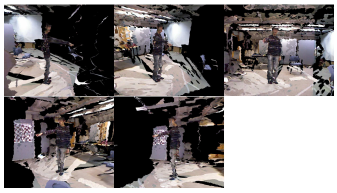

図 7: 歪み補正後の指差しの三次元復元結果の正面 5 方向からの描画

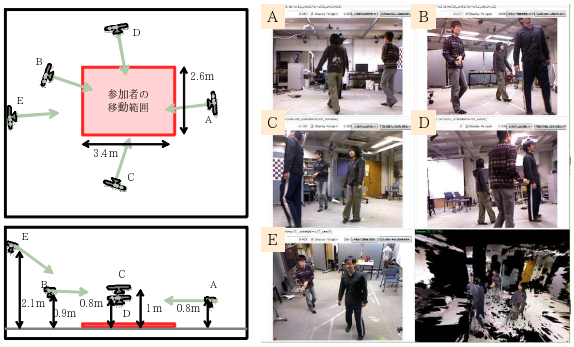

図 8: 三人が立ち話をしている場面を収録した際の Kinect の配置とその映像

影した (図 8)．会話環境内で人の動く範囲はおおよそ 3m 四方程度，収録した時間としては約 80 秒である．

まず，図 8 のなかで，斜め上から撮影している Kinect(E) から得られる骨格座標を正解データとして，他の Kinect から得られる骨格座標を統合したものと比較した．会話参加者は，必ずしも Kinect の方を向いているわけではないので，両肩の骨格座標のカメラ水平距離が，平均以上のフレームのみを正面を向いているものとして評価対象とした．

表 1: 正解データの座標と，統合後の座標との距離 [mm] の平均と標準偏差の比較

		平均	標準偏差
被験者 1	本手法による統合	97.4	51.3
	平均による統合	111.7	54.3
被験者 2	本手法による統合	104.0	53.9
	平均による統合	110.6	52.1
被験者 3	本手法による統合	117.5	59.4
	平均による統合	124.3	62.9

正解データと，統合した骨格データの比較結果を表 1 に示す．それぞれの値は，各フレームの各骨格点について，正解データにおける座標と統合した結果による座標との距離の平均を表す．

本手法による統合は，単純な平均による手法と比べると，全体的に推定精度が上昇していると考えられる．課題としては，非接触のモーションキャプチャデバイスではなく，接触型のより精度の高いデバイスを用いてそれを正解データとし，その結果と今回の推定を比べることでより信頼性の高い評価を行うことが挙げられる．

6.2.3 会話シーンの三次元記録と会話環境の三次元モデルの統合

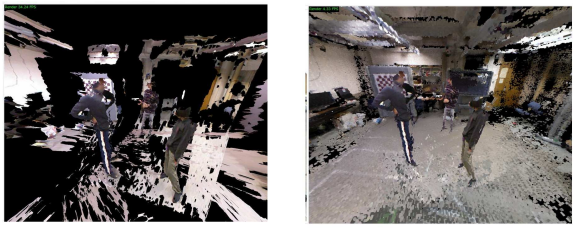

図 9: 会話シーンの復元時の背景 (左) と，会話環境の三次元復元と会話シーンの人領域の三次元復元とを重畳させて表示したもの (右)

図 9 は会話シーンの三次元復元と会話環境の三次元復元を統合して表示させたものである．図の左は比較用の会話シーンの Kinect からの映像のみを用いたものある．右が統合して表示したものであり，会話シーンは人領域のみを三次元復元し，背景に会話環境の復元で得られた三次元モデルを描画している．会話シーン撮影時の環境側と同じ位置に復元した環境の三次元モデルが描画されている．また，三次元モデルとして人と物の位置関係が分かるようになっている．

7. おわりに

本論文では，会話データを映像としてさまざまな位置から再現して会話分析者が閲覧できるようにするための，複数台の Kinect を用いて会話シーンと会話環境を記録し，統合や補正を行って記録データを三次元的に再現するシステムを構築した．

システムの今後の課題としては，歪み補正や環境の三次元モデル復元に対してより高速な手法を用いることで，収録から会話分析へと進めるためにかかるデータ処理の時間を削減すること，シーンや環境を復元したモデルのデータ構造を変更し閲覧システムとして使いやすくすることが挙げられる．

参考文献

[Maimone 11] Maimone, A. and Fuchs, H.: Encumbrance-Free Telepresence System with Real-Time 3D Capture and Display using Commodity Depth Cameras, in *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, pp. 137–146IEEE (2011)

[McCowan 05] McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., et al.: The AMI meeting corpus, in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, Vol. 88 (2005)

[Moeslund 01] Moeslund, T. and Granum, E.: A survey of computer vision-based human motion capture, *Computer Vision and Image Understanding*, Vol. 81, No. 3, pp. 231–268 (2001)